

国立国会図書館デジタル化資料データ（画像・テキスト）の使い道 90分アイデアソンの進め方

16時～16時15分 アイデアソン説明【15分】

アイデアソンのインプットとして、当館のOCR関連事業、オープンデータ及び実験サービスを紹介

16時15分～17時 ブレイクアウトセッション【45分】

【アイデアソン参加者】

3班×5-6名に分かれ、各ブレイクアウトルームで議論を行う

【一般参加者（聴講者）】

- ・アイデアソン参加者の議論の様子は、オンラインホワイトボードの共有により確認できる
- ・モレーテタが最初に導入的なプレゼンを行い、次にホワイトボードを見ながら、各ブレイクアウトルームの議論の様子を視聴者向けに紹介＆コメントをする

17時～17時30分 アイデアソン参加者各班の発表、総括【30分】

各班から議論の結果を発表し（発表4分+質疑4分）、モレーテタが講評及び全体の総括を行う

国立国会図書館デジタル化資料データ (画像・テキスト) の使い道：90分アイデアソン インプット資料

国立国会図書館電子情報部電子情報企画課
次世代システム開発研究室 青池亨

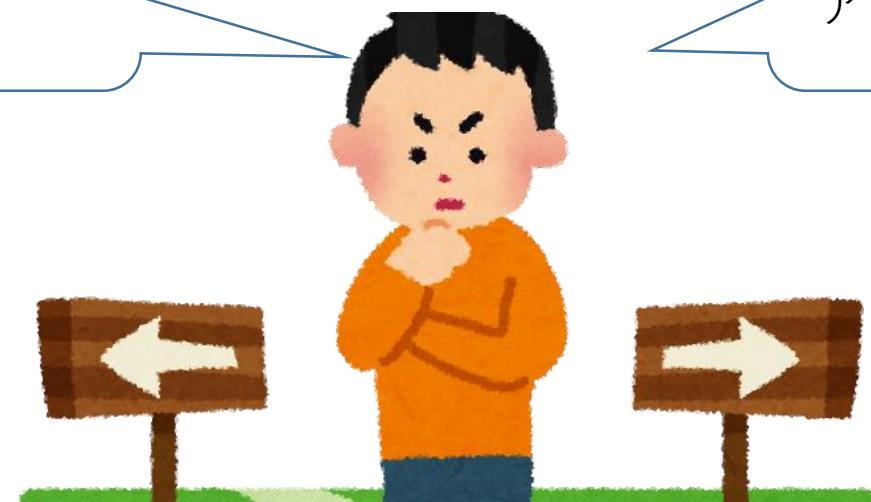
この資料の見方



データを検索しながらチラ見して
アイデアを膨らませたい

まとめたデータをダウンロードして手元で
アイデアを膨らませたい

【サービス】についている
スライドをご覧ください



【データセット】についている
スライドをご覧ください
(※データセット①を除く)

本日のアイデアソンのために詳しく説明するデータセット

国立国会図書館（NDL）では、令和2年度補正予算（第3号）により、令和3年度に2つの事業を実施しました

（参考：https://lab.ndl.go.jp/data_set/ocr/）

事業1. 当館がこれまで提供していたほぼ全てのデジタル化資料（約247万点）を当館資料用に特別に性能改善したOCRでテキスト化する事業

事業2. 今後当館がデジタル化する資料に対してテキスト化処理を行うOCR処理プログラムを研究開発する事業

これらの事業によって、新たに提供可能になったデータセットを主に取り上げます

NDLが提供しているその他のオープンデータセット (参考情報)

「書誌データ」「典拠データ」

「国会会議録テキストデータ・画像データ」

「全国の図書館や公文書館によるレファレンス事例」

といったものがあります

- <https://www.ndl.go.jp/jp/dlib/standards/opendataset/index.html>
- https://lab.ndl.go.jp/data_set/dataset/

本日は、OCR事業の成果物であるデータセットを中心にお話しします
が、是非これらのデータ資源もご活用ください！

【データセット①】 OCRテキストデータ

どんなデータセット？

- ・事業1で作成された、247万点分の全文テキストデータ
- ・座標情報や縦書き・横書きといった情報も含まれる
- ・自動で読み取られた結果なので、一定の割合で誤りがある

データセットのサンプル

<https://lab.ndl.go.jp/dataset/joss2022/OCRtxtdatasample.zip>

本日の「主菜」

とはいっても、分量が膨大なので、生データをぽんと渡されても、一日では把握しきれないと思います

中身をつまみ食いでできる実験サービスを後ほどご紹介します

【データセット①】 OCRテキストデータ

OCRテキストデータが持つ情報（テキストボックス単位の情報）

Key	Description
words	テキストボックス情報のリスト
id	テキストボックスの順序
boundingBox	テキストボックスの座標位置
isVertical	縦書きか横書きか
text	文字列単位に区切られたテキストデータ 縦書きの場合は文字が正立する向きで上から、横書きの場合は左から順に出力される
confidence	文字認識の信頼度 0-1の範囲を取り、大きいほど信頼度が高い
isTextline	本文か、それ以外か（文字サイズによる判定）
isRTL	右横書きか、左横書きか

【データセット①】 OCRテキストデータ

OCRテキストデータが持つ情報（1行単位の情報）

Key	Description
lines	行ボックス情報のリスト
id	行ボックスの順序
boundingBox	行ボックスの座標位置
wordIDs	行に含まれるテキストボックスのidリスト
estimatedLanguage	言語の推定結果 ja: 日本語, ko: 韓国語, ta: 台湾語, en: 英語 * テキストボックスが1つも検出されなかつた場合、“NULL”(文字列)を出力

【データセット②】 OCR処理プログラム (NDLOCR)

※便宜上データセットと書きましたが、プログラムです

https://github.com/ndl-lab/ndloctr_cli

どんなプログラム？

- ・デジタル化資料に記述された文字列を読み取ってテキスト化するプログラム
- ・そのまま使うにはシステムの知識が必要だが、東京大学の中村覚助教らにより、簡単に使えるチュートリアルが用意されている

<https://zenn.dev/nakamura196/articles/b6712981af3384>

【データセット③】 OCR学習用データセット

どんな人に向いている？

自分で機械学習を学んでOCR処理プログラムを作りたい方や、
今利用しているOCRの性能を評価したい方

どんなデータセット？

- 資料画像
- 資料画像の内部に書かれた正解テキスト情報
- 一部のデータにはレイアウト情報（キャプション、タイトル、著者名等）も入っている

【データセット③】 OCR学習用データセット

- <https://github.com/ndl-lab/pdmocrdataset-part1>

OCRテキスト化事業の性能改善を目的として、当館の保有するデジタル化資料から作成したOCR学習用途の機械学習データセットのうち、著作権保護期間の満了した資料から作成されたデータセット（2,713画像分）

- <https://github.com/ndl-lab/pdmocrdataset-part2>

NDLOCRの学習を目的として当館の提供するデジタル化資料から作成したOCR学習用途の機械学習データセットのうち、著作権保護期間の満了した資料から作成されたデータセット（3,997画像分）

【データセット④】

レイアウトデータセット・図版タグデータセット

著作権保護期間満了資料から、NDLラボが作成した画像のデータセット

NDL-DocL（資料画像レイアウトデータセット）

<https://github.com/ndl-lab/layout-dataset>

→古典籍資料と明治以降刊行資料についてそれぞれ作成

NDL-ImageLabel（ラベル付画像データセット）

<https://github.com/ndl-lab/imagetagdataset>

→自動で切り出された図版を、写真の種類やイラストの種類で分類してタグ付けしたデータセット

【サービス①】 次世代デジタルライブラリー

<https://lab.ndl.go.jp/dl/>

国立国会図書館デジタルコレクションで提供しているデジタル化資料の中から、著作権の保護期間が満了した図書及び古典籍資料全部（約33万6千点）が検索可能（全文検索は図書のみ約28万点）な実験サービス

何ができるサービス？

- ・デジタル化資料の「中身」を全文検索＆画像検索できる
- ・資料ごとにOCRテキストデータをダウンロードできる

【サービス①】 次世代デジタルライブラリー

全文から検索する

クリエイティブ 

本文のみを検索する 検索結果に図版を表示しない

詳細検索

2件見つかりました

1 < > 表示件数 20件 ソート 一致度順

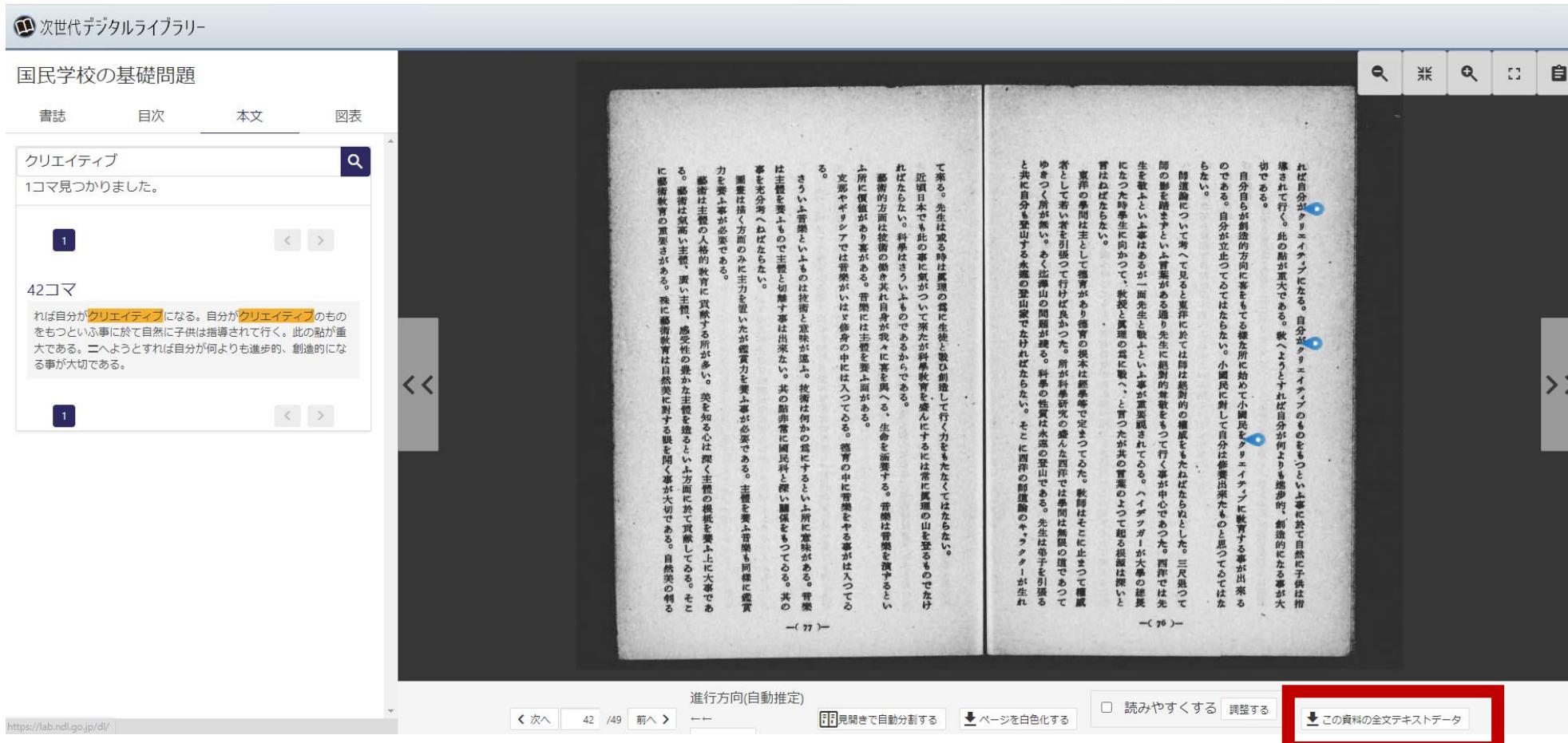
 国民学校の基礎問題
木村素衛 [述]諏訪郡永明国民学校講読会 1941

れば自分がクリエイティブになる。自分がクリエイティブのものをもつといふ事に於て自然に子供は指導されて行く。此の點が重大である。二へようとすれば自分が何よりも進歩的、創造的になる事が大切である。

自分自らが創造的方向に喜をもてる様な所に始めて小國民をクリエイティブに育する事が出来るのである。自分が立止つてゐてはならない。小國民に對して自分は修業出来たものと思つてゐてはならない。

<https://lab.ndl.go.jp/dl/fulltext?keyword=%E3%82%AD%E3%82%A2%E3%82%A4%E3%83%83%E3%83%BC&searchfield=contentonly>

【サービス①】 次世代デジタルライブラリー



次世代デジタルライブラリー

国民学校の基礎問題

書誌 目次 本文 図表

クリエイティブ

1コマ見つかりました。

42コマ

42 / 49

1

1

22

26

進行方向(自動推定)

次へ 42 / 49 前へ

見開きで自動分割する ページを白色化する

□ 読みやすくする 調整する

この資料の全文テキストデータ

れば自分がクリエイティブになる。自分がクリエイティブのものももつといふ事に於て自然に子供は指導され行く。此の點が重大である。教へようとするが自分よりも進歩的、創造的になる事が大切である。自分が創造的方向に事をもてる様な所に始めて小国民はクリエイティブに教育する事が出来るのである。自分が立止つてゐるはならない。小国民に對して自分は絶対的出来た。と思つてゐてはならない。

開道論について著してゐると西洋には師は絶対的の権力をもたねばならないとした。三尺進つて師の影を踏まずといふ事がある通り先生に絶対的尊敬をもつて行く事が中心であつた。西洋では先生を教ふといふ事はあるが一回先生と教ふといふ事が重要視されてゐる。ハイダッガーが大學の経営になつた時学生に向かつて、教授と眞理の爲に教へ、と言つたが其の言葉のよつて起る根柢は深いと言はねばならない。

西洋の學問は主として德育があり德育の根本は経済等で定めつてゐた。德育はそこには止まつて德育者として若い者を引摺つて行けば良かった。所が科學研究の進んだ西洋では學問は無限の道であつてゆきつく所が無い。多く迄得山の問題が残る。科學の性質は永遠の登山である。先生は弟子を引摺ると共に自分も登山する主義の登山家でなければならない。そこに西洋の開道論のキラクターが生れる。

て来る。先生は或る時は眞理の爲に生徒と戦ひ創造して行く力をもたなくてはならない。眞理の本でも此の事に気がついて來たが科學教育を盛んにするには常に眞理の山を登るものでなければならない。科學はさうかふものであるからである。

藝術的方面は技術の能く其自身が我々に參與へる、生命を發揮する。音樂は音樂を演ずるといふ所に藝術があり喜がある。音樂には主體を裏面がある。

支那やギリシアでは音樂がはゞ修身の中にひつてゐる。德育の中に音樂をやる事がはひつてゐる。

さういふ音樂といふものは技術と意味が運ぶ。技術は何かの爲にくるといふ所に意味がある。音樂は主體を要ふもので主體と切離す事は出来ない。其の點非常に國民的と深い關係をもつてゐる。其の事を充分考へねばならない。

圓形はよく方圓のみに主力を置いていたが圓實力を要す事が必要である。主體を要す音樂も同様に圓實力を要す事が必要である。

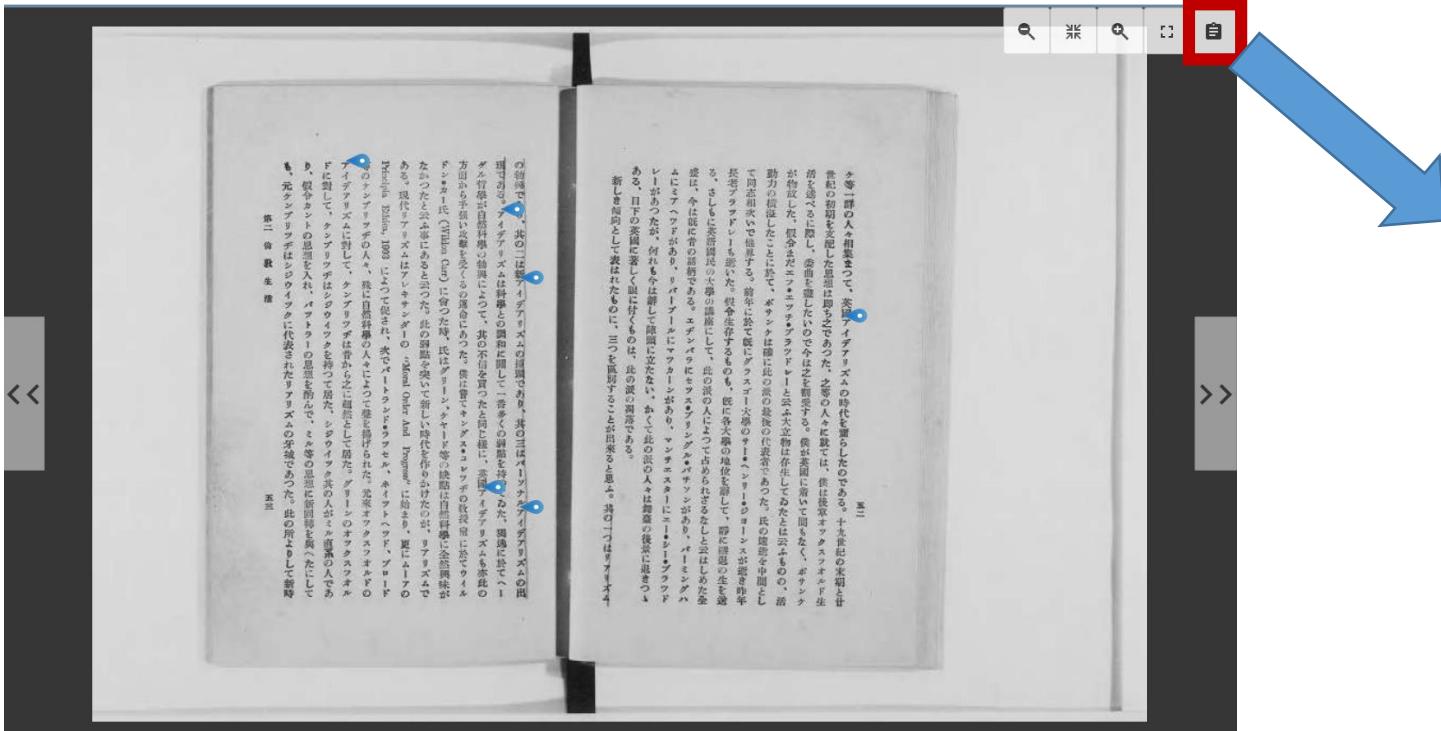
藝術は主體の人格教育に貢献する所が多い。美を知る心は深く主體の根柢を要すに大事である。藝術は高い主體、廣い主體、感受性の豊かな主體を造るといふ方面に於て貢献してゐる。そこには藝術教育の重要さがある。殊に藝術教育は自然美に對する眼を開く事が大切である。自然美の判る

<https://lab.ndl.go.jp/dl/book/1437246?keyword=クリエイティブ&page=42>

OCRテキストデータのダウンロードボタン

【サービス①】 次世代デジタルライブラリー

- OCRテキスト表示機能・コピペ機能



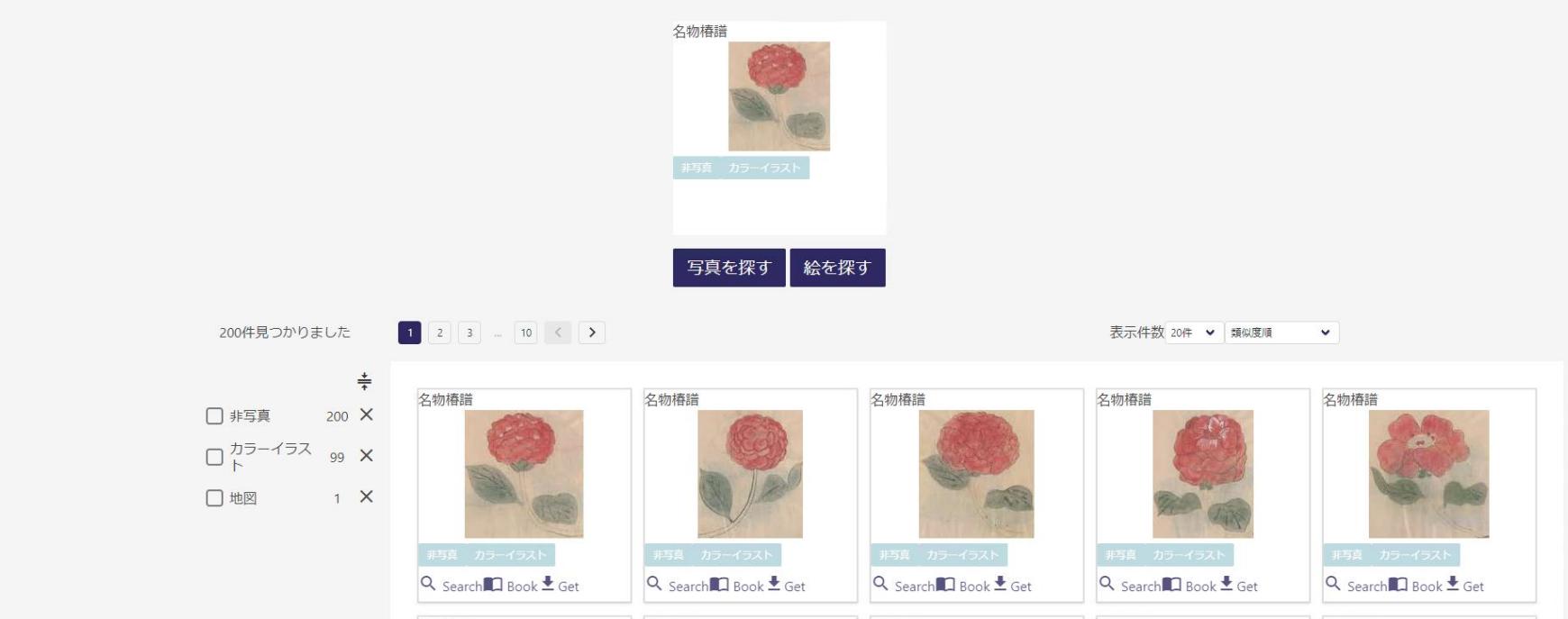
の勃興であり、其の二は新アイデアリズムの擡頭であり、其の三はパーソナルアイデアリズムの出現である。アイデアリズムは科學との調和に関して一番多くの弱點を持つてゐた、獨逸に於てヘーゲル哲學が自然科學の勃興によって、其の不信を買つたと同じ様に、英國アイデアリズムも亦此の方面から手強い攻撃を受けるの運命にあつた。僕は嘗てキングス・コレッヂの二授室に於てウイルドン・カーハ氏(WildenCH)に會つた時、氏はグリーン、ケヤード等の缺點は自然科學に全然興味がなかつたと云ふ事にあると云つた。此の弱點を突いて新しい時代を作りかけたのが、アリksamenderの“MoralOrderAndProgress”に始まり、更にムーアのPrincipiaEthica,1903によつて促され、次でパートランド・ラツゼル、ホイツトヘッド、ブロード等のケンブリッヂの人々、殊に自然科學の人々によつて聲を揚げられた。元來オックスフォルドのアイデアリズムに對して、ケンブリッヂは昔から之に超然として居た。グリーンのオックスフォルドに對して、ケンブリッヂはシジウイツクを持って居た、シジウイツク其の人がミル直系の人であり、假令カントの思想を入れ、バツトラーの思想を酌んで、ミル等の思想に新回轉を與へたにして元ケンブリッヂはシジウイツクに代表されたアリズムの牙城であつた。此の所以よりして第2倫敦生活五三五ニケ等一群の人々相集まつて、英國アイデアリズムの時代を齎したのである。十九世紀の末期と廿世紀の初期に支配した思想は即ち之であつた、之等の人々に就ては、僕は後章オックスフォルド生活を述べるに際し、委曲を盡したいので今は之を割愛する。僕が英國に於いて間もなく、ボサンケが物故した、假令まだエフ・エッソ・ブラッドレーと云ふ大立物は生存してゐたと云ふものの、活動力の横溢したことにして、ボサンケは確に此の派の最後の代表者であつた。氏の遠逝を中間として同志相次いで他界する。前に於て既にグラスゴー大學のサー・ヘンリージョーン斯が逝き昨年長老ブラッドレーも逝いた。假令生存するものも、既に各大學の地位を辭して、靜に隠退の生を送る、さしもに英語國民の大學生の講座にして、此の派の人によつて占められざるなしと云はしめた全盛は、今は既に昔の話柄である。エシバンラにセツ・ブリンクルバチソンがあり、バーミンガムにミアヘッドがあり、リバーブールにマツカーンがあり、マンチエスターにエー・シー・ブラッドレーがあつたが、何れも今は辭して陣頭に立たない、かくて此の派の人々は舞臺の後景に退きつゝある、目前の英國に著しく眼に付くものは、此の派の洞窟である。新しき傾向として表はれたものに、三つを區別することが出来ると思ふ。其の一つはアリズム

コピー 範囲選択 開じる

<https://lab.ndl.go.jp/dl/book/980819?keyword=%E3%82%AD%E3%82%BB&page=36>

【サービス①】 次世代デジタルライブラリー

NDLラボで開発した機械学習モデルにより、資料中の図版について、類似図版の検索も可能



The screenshot shows a search interface for a red rose illustration. At the top, there is a search result for a red rose with the label '名物椿譜' (Monogatare of Camellias). Below it are two buttons: '写真を探す' (Search for photo) and '絵を探す' (Search for drawing). The search results page displays 200 items found, with a navigation bar showing pages 1 through 10. On the left, there is a sidebar with filters: '非写真' (200), 'カラーイラスト' (99), and '地図' (1). The main results area shows five preview cards, each with a red rose illustration, a 'Search' button, a 'Book' button, and a 'Get' button. The cards are labeled '名物椿譜' and '非写真 カラーイラスト'.

https://lab.ndl.go.jp/dl/illust/search?image=2535667_5_1

【サービス②】 NDL Ngram Viewer

<https://lab.ndl.go.jp/ngramviewer/>

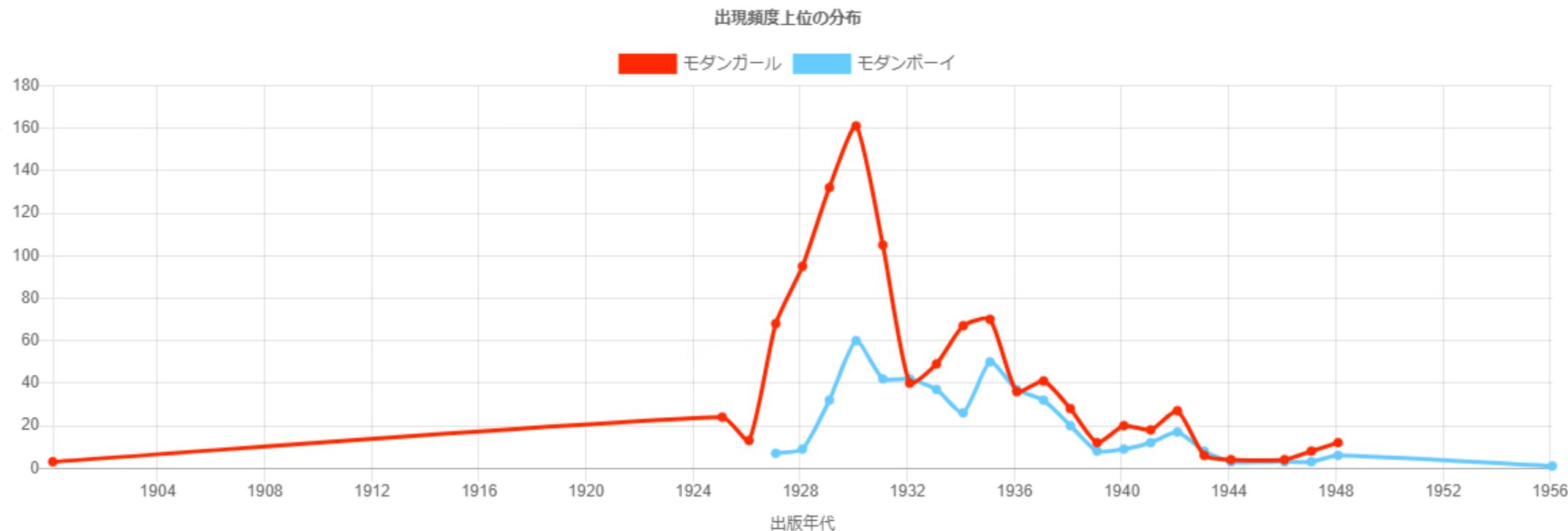
国立国会図書館デジタルコレクションでインターネット公開している資料のうち、著作権保護期間が満了した図書資料約28万点（次世代デジタルライブラリーの全文検索対象と同じ資料群）のOCRテキストデータから集計した、約8.3億種類の単語及びフレーズの出現頻度を、出版年代ごとに可視化できるツール

何ができるサービス？

- OCRテキストデータ中に存在する単語や語用法の出現頻度の可視化
- 正規表現による周辺キーワードの探索

【サービス②】 NDL Ngram Viewer

「モダンボーイ」と「モダンガール」の出現頻度を比較

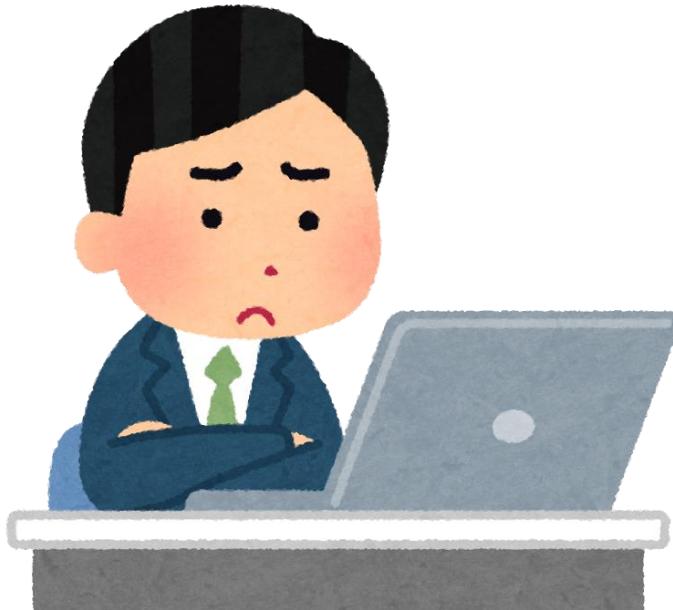


<https://lab.ndl.go.jp/ngramviewer/?keyword=%E3%83%A2%E3%83%80%E3%83%B3%E3%82%AC%E3%83%BC%E3%83%AB%2F%E3%83%A2%E3%83%80%E3%83%B3%E3%83%9C%E3%83%BC%E3%82%A4&size=100&from=0>

【サービス②】 NDL Ngram Viewer

例えば「関ヶ原の合戦」について調べたいとき

「関ヶ原」？
「関ヶ原」？
それとも「関が原」？



「関ヶ原の合戦」？
「関ヶ原合戦」？
それともひょっとして
「関ヶ原の戦い」？

【サービス②】 NDL Ngram Viewer

→ 「関.原.*戦い？」で正規表現検索！

<https://lab.ndl.go.jp/ngramviewer/?keyword=%E9%96%A2.%E5%8E%9F.%2a%E6%88%A6%E3%81%84%3F&size=100&from=0>

キーワード	総出現頻度	
関ヶ原の戦	2010	次世代デジタルライブラリーで検索
関ヶ原合戦	1258	次世代デジタルライブラリーで検索
関ヶ原の戦	967	次世代デジタルライブラリーで検索
関ヶ原合戦	824	次世代デジタルライブラリーで検索
関ヶ原の一戦	626	次世代デジタルライブラリーで検索
関ヶ原の合戦	466	次世代デジタルライブラリーで検索
関ヶ原戦	431	次世代デジタルライブラリーで検索
関が原の戦	394	次世代デジタルライブラリーで検索
関ヶ原ノ戦	352	次世代デジタルライブラリーで検索
関ヶ原大戦	346	次世代デジタルライブラリーで検索
関ヶ原の合戦	284	次世代デジタルライブラリーで検索
関ヶ原の大戦	201	次世代デジタルライブラリーで検索
関ヶ原戦	181	次世代デジタルライブラリーで検索

【サービス②】

NDL Ngram Viewer

演算子	クエリ例	説明
.	<u>修.者</u>	任意の1文字を表す
*	<u>日*精進</u>	直前の表現が0個以上あることを表す
+	<u>郵便局+長</u>	直前の表現が1個以上あることを表す
?	<u>巡回部?長</u>	直前の表現が0個か1個あることを表す
{,}	<u>私を.{2,3}にする</u>	直前の文字の繰り返し回数の範囲を指定する。例1:{2,4}:2回以上4回以下 例2:{3,}:3回以上 例3:{,5}:5回以下
[…]	<u>大正[ーーーン]{6,}</u>	ブラケット内の1文字を表す。「-」で範囲を、「^」で否定を表す。例1[abc]:a,b,cのうち1文字 例2[アーーーン]：アからンまでの1文字（カタカナいずれか1文字） 例3[^アーーーン]：アからンまで以外の1文字（つまりカタカナ以外の1文字）
(…)	<u>春の海ひねもす(のたり)*</u>	かっこで囲んだ範囲のグループを形成し、単一の文字として扱う。（他の演算子と組み合わせて用いる）
	<u>ご(機嫌 きげん)よう</u>	左辺または右辺の最長のパターンにマッチすることを表す

Q.著作権保護期間の存続している資料のデータは使えないの？

A.今日のところは残念ながら提供できませんが、条件を満たせばお渡しできます！

- 著作権保護期間の存続している資料から作成したOCR学習用データセットについては、原資料の著作権保護の観点からGitHub上に公開することはできません。
- これらのデータに関しては、当館との協議のうえで著作権法上認められた範囲内の利用（著作権法第30条の4の規定による機械学習目的など）に限り、当館と書面を取り交わした上で提供することが可能です。

Q. OCRテキストデータやNDLOCRの性能はどのくらい？

A.事業の報告資料を公開していますのでご覧ください！

OCRテキストデータの報告資料

https://lab.ndl.go.jp/data_set/ocr/r3_line/

NDLOCRの報告資料

https://lab.ndl.go.jp/data_set/ocr/r3_morpho/

Q. 提供されているデータセットを利用して商用サービスを開発しても大丈夫？

A. 問題ありません、大丈夫です！

データセットのライセンスはいずれもCC BY 4.0で提供しています

NDLOC Rについても、新規開発部分はCC BY 4.0、既存のライブラリ等を利用している部分については寛容型オープンライセンスのものを採用しているため、商用非商用を問わず自由な改変、利用が可能です。

アイデアソン参加者の皆さんへ

- 同じ班の誰かが良い知恵を持っていらっしゃるかもしれません、是非、「こんなことできたら嬉しいな」を気軽に投げかけてください（特に前半のアイデア出しの段階）
- 専門知識をお持ちの方は、「こうしたらうまくいくよ・こんな事例があるよ」を積極的に共有して頂けると大変ありがとうございます
- 今後データセットを使って共同研究等を進めていく仲間づくりの場としていただいても大歓迎です

アイデア出しのコツ

- ・ プレイズ・ファースト (Praise First)

「先に褒めよ」

アイデアの良い所を褒める。批判はあと
→アイデアを育てる

「プレイズ・ファースト」の本質は「良点発見」（ブレインストーミングのルール+α）
石井力重の活動報告.<http://ishiirikie.jpn.org/article/67402619.html> (2022-06-08 accessed)

一般参加者の皆さんへ

- ・ブレイクアウトセッション中は、ぜひチャット機能を活用して活発に質問やコメントをしてください
- ・質疑の時間では、アイデアの実現につながるようなご意見を歓迎します
- ・もしよろしければ、SNS等にも書き込んで頂ければと思います
ハッシュタグ： **#JOSS2022 #90分アイデアソン**

全ての参加者の皆さんへ

- ・今回のアイデアソンによって、集まつたいろいろなバックグラウンドの皆さんの間で交流が生まれ、オープンデータから新しいサービスが育っていくきっかけになることを願っています
- ・一人では思いつかない素敵なアイデアを議論の中に発見して頂ければと思います
- ・アイデアソンが終わった後が本番……になると嬉しいです。イベント終了後も気兼ねなく、是非是非お問い合わせください

さいごに

アンケートにご協力お願いします

<https://enquete.ndl.go.jp/249638>

お問い合わせは、NDLラボ lab@ndl.go.jp まで